DOCUMENT RESUME

ED 064 301 TM 001 419

AUTHOR Modu, Christopher C.
TITLE The Effectiveness of an Essay Section in the American History and Social Studies Test.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO RB-72-5
PUB DATE Feb 72
NOTE 49p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Academic Achievement; *American History; *Essay Tests; *Objective Tests; Psychometrics; *Social Studies; *Test Reliability; Test Results

ABSTRACT

The contribution of a 20-minute essay question, given as part of the one-hour achievement test in American History and Social Studies, to the pool of information available on a candidate from an all-objective examination of the College Board Admissions Testing Program is presented in this report. The study limits itself to a consideration of the psychometric issues only. It does not deal with other important issues related to the educational impact of the essay or to the implications of the cost of introducing and maintaining a reliable essay section in the American History and Social Studies test. The results of the study led to the conclusion that the additional information gained by the introduction of an essay section is minimal. Moreover, the correlation of the essay with various subtests of the American History and Social Studies objective test suggests that the obtained essay scores may have reflected more of the factual content of the answers than the other skills. (Author)

COLLEGE ENTRANCE EXAMINATION BOARD
RESEARCH AND DEVELOPMENT REPORTS
RDR-71-72, NO. 6

RESEARCH BULLETIN
RB-72-5 FEBRUARY 1972

The Effectiveness of an Essay Section in the American History and Social Studies Test

Christopher C. Modu

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY
BERKELEY, CALIFORNIA

THE EFFECTIVENESS OF AN ESSAY SECTION IN THE AMERICAN HISTORY AND SOCIAL STUDIES TEST

Christopher C. Modu

Educational Testing Service

Abstract

The contribution of a 20-minute essay question, given as part of the one-hour achievement test in American History and Social Studies, to the pool of information available on a candidate from an all-objective examination of the College Board Admissions Testing Program, is presented in this report. The study limits itself to a consideration of the psychometric issues only. It does not deal with other important issues related to the educational impact of the essay or to the implications of the cost of introducing and maintaining a reliable essay section in the American History and Social Studies test. The results of the study led to the conclusion that the additional information gained by the introduction of an essay section is minimal. Moreover, the correlation of the essay with various subtests of the American History and Social Studies objective test suggests that the obtained essay scores may have reflected more of the factual content of the answers than the other skills.

THE EFFECTIVENESS OF AN ESSAY SECTION IN THE AMERICAN HISTORY AND SOCIAL STUDIES TEST[1]

The Problem and Its Background

The purpose of this study is to investigate the contribution of a 20-minute essay question, given as part of the one-hour achievement test in American History and Social Studies, to the pool of information obtainable on a candidate from the all-objective examinations of the College Entrance Examination Board's Admissions Testing Program.

Prior to the study, the achievement test in American History and Social Studies consisted of a one-hour objective test of about 100 items. However, in response to pressures from advocates of an essay-type examination who have often expressed the conviction that the distinctive qualities of an essay test would provide unique information about candidates, it was decided to investigate the effectiveness of an American History and Social Studies essay by including a 20-minute essay in the operational section of the March 1971 administration of the Admissions Testing Program.

The investigation had to face the question of whether a 20-minute essay compared favorably with other types of questions, whether it was reliable enough to be able to compete in the prediction of a criterion performance in the subject, whether it would contribute significantly to the discrimination of candidates according to their ability levels, and whether or not the essay was more effective than the objective test for candidates at various achievement levels.

Certain advocates of the essay had argued at some of the Committee meetings which the investigator attended that regardless of how little the essay might contribute to the rank-ordering of candidates, the mere fact of its introduction in the college admissions program would have a salutary effect on the teaching of the subject in the schools and on how the candidates prepare for it.

Members of the Committee also argued that essay questions would not only require a candidate to supply and organize his own data but also force him to attempt a formulation of his explanations in what, hopefully, would be a clear-cut and logical argument; that many students go to colleges and universities unprepared to deal with essay questions which they would invariably face in college history courses and that the inclusion of an essay in College Board achievement tests would intensify the emphasis on essay tests in high schools.

On the other hand, arguments against the use of essay tests emphasized the premium placed on literary fluency rather than mastery of a field; that an essay examination had little to do with the skills of the historian since it did not allow time for research and thoroughness in the preparation of conclusions with the result that a student who could not write under pressure would be unduly penalized.

Although the argument for the use of essay tests to influence secondary school curricula seems like a sound peg on which to hang the demand for changes in examining methods, it could be equally argued that since examinations are tools to be used rather than goals to be aimed at, justification for the use of essay tests should be based on the validity of a combined essay and objective

test relative to that of an all-objective test. However, since operational and cost considerations are involved in a final decision on the introduction of an essay component, the problem is one of determining whether achievement in the subject as taught in schools is better assessed by a combination of both types of tests than by an objective test alone. Posed in this manner, the problem statement puts the burden of demonstrating its own usefulness on the essay, placing it, in effect, in a defensive position. Clearly, either type of test has certain advantages and disadvantages. However, the advantages claimed for the essay test must outweigh the economy, the faster and more reliable marking, the wider subject-matter coverage, the gain in reliability and validity and the better spread of marks from objective tests to justify an attempt to circumvent the danger inherent in the use of solely objective tests by the introduction of an essay section.

There is also the problem of whether the sample of a student's ability in a 20-minute essay is so small and unrepresentative as to make little difference in the findings. However, this study does not address itself specifically to that question. The decision having been taken to give a 20-minute essay part of the one-hour achievement test, the task here is to determine what difference it then made.

It had been demonstrated (Godshalk, Swineford, & Coffman, 1966) that 20-minute essay scores based on five independent readings contributed uniquely to the validity of a one-hour English Composition Test for predicting a reliable essay criterion. It was also shown in the same study that the unique validity of a 20-minute essay could be obtained from two or three readings instead of four or five, particularly if a four-point scale were substituted for the three-

point scale used for the five independent readings. Huddleston (1954) also compared the validities of the Verbal Sections of the Scholastic Aptitude Test and of an objective editing test with those for essay ratings based on two paragraphs composed by the students. Using as criteria English grades and systematic ratings made by teacners who knew the students well, Huddleston concluded that measurable "ability to write" was no more than verbal ability as measured by the SAT-verbal test.

The studies cited above concentrated on the unique contribution of essay scores in English Composition to a measure of writing ability. The present study, on the other hand, was not merely an attempt to confirm the unique validity of an essay test by using a cognate set of subjects in a different field but sought to determine the contribution of a 20-minute essay in American History and Social Studies to a battery of objective tests of the Admissions Testing Program, comprising subjects that are not as closely related to one another as SAT-verbal and English Composition. Based, therefore, on the assumption that other tests of the College Board Admissions Testing Program sample areas of mental abilities which overlap with those covered by the American History and Social Studies test, the problem was defined and explored in such a way as to determine the relative contribution of the essay to the candidates' overall performance on the College Board admissions examinations as a basis for judging its effectiveness.

It was recognized from the start that what was needed for a validity study of the American History and Social Studies essay test was a logically sound criterion of achievement in the subject--a criterion devoid of any extraneous outcomes of classroom instruction. However, no plans were made for

any elaborate collection of criterion data designed to measure the desirable or ideal objectives of an American History course nor would the operational aspects of the Admissions Testing Program allow for an experimental manipulation of the variables with control and experimental groups; thus, the investigator had to settle for alternative, if less powerful, methods of finding an answer to the question under study.

Brief Statement of Results

The results of the study demonstrate that the 20-minute essay section in American History and Social Studies measured a different psychological trait from that of the 40-minute objective section, and that the essay test was primarily a measure of achievement in the subject rather than one of writing ability.

Insofar as the unique contribution of the essay to the battery of tests (made up of the 40-minute objective section in American History and Social Studies, SAT-verbal, SAT-mathematical and English Composition Test) is concerned, its effectiveness, as judged by the additional information gained from its scores, is quite minimal. This was demonstrated by the rank-order of the study sample on the composite score on the battery of four tests (excluding the essay) which hardly made a noticeable change when the essay scores were included. Only when its weighting was increased from about one-twelfth to three-tenths (where weights on the other tests of the battery were such that the total weight is unity) does the essay begin to make an appreciable change in the candidates' rank-order. Furthermore, the proportion of true-score variance in the essay which is both reliable and completely free of overlap with

any of the other tests in the battery was found to be only about .16. It is this proportion that must be judged and in the author's opinion it is not important. This conclusion is further supported by the factor analysis in which the essay accounted for only 2.7% of the total variance in the common factor space of the subscores into which the four tests used for the study were subdivided.

Moreover, the widely held assumption that objective tests measure less complex mental processes than those required for essay tests, such as originality of production and organization of ideas, was not supported by the essay test used for this study. Instead, the essay scores correlated higher with the subscores of the objective section which reflected factual information than with those which measured the ability to apply hypotheses or to synthesize and judge the value of data for a given purpose.

The Design of the Study

The plan of the study called for the administration of a mixed essay and objective test in American History and Social Studies (often abbreviated from here on as AHSS) at the March 1971 examinations of the College Board Admissions Testing Program. The various analyses were based on a representative sample of candidates who offered, in addition to the AHSS test, the verbal and mathematical sections of the Scholastic Aptitude test as well as the English Composition Test. The form of the English Composition Test used at the March, 1971 administration contained 95 multiple-choice items of appropriate English usage, error recognition and "construction shift" items. The "construction shift" item-type requires the candidates to make a change in one part of a

sentence based on a revision that he is given in another part of that sentence. An entirely objective English Composition Test is offered interchangeably with an English Composition Test that requires the writing of an essay.

The English Composition Test (ECT) was selected both because of the likelihood of its having a much higher correlation with AHSS than with any of the other achievement tests (Biology, Chemistry, Physics and Mathematics) offered at this administration and the fact that slightly over 86% of the candidates who took AHSS also sat for English Composition. Therefore, the ECT score would be one that admissions officers could use for most candidates offering AHSS scores. Another reason for including ECT was to examine both by correlational and other techniques whether the AHSS essay scores were essentially different from the scores obtained on a test of writing skills.

The analyses were carried out in five stages as follows:

(1) Correlation methods in which the interrelations among the variables and the unique contribution of the AHSS essay were explored.

(2) Multiple regression analyses for investigating the extent to which the scores in AHSS essay could be predicted from a combination of the scores in SAT-verbal, SAT-math, AHSS objective and English Composition Test, the underlying assumption being that a successful prediction of a candidate's AHSS score from the scores on the other tests would render the essay redundant. The regression analyses were carried out both with the total sample and again with four different categories of the sample

divided according to their combined aptitude test scores in order to verify whether the essay score is equally predictable at each level of aptitude score.

(3) Factor analysis of the subscores into which each of the four tests used for the study have been subdivided according to an assumed hierarchical order of the complexity of the mental processes involved in answering the items in each subgroup. The AHSS essay test was treated as a single subtest in the factor analysis in order to determine (a) whether it could be identified as a unique factor and (b) the proportion of variance due to it in the common factor space.

(4) Discriminant function analysis in which the contribution of each of several independent variables (including AHSS essay) for distinguishing (or "discriminating") between subgroups of the total sample divided according to their ability levels on a specified criterion variable was determined.

(5) An exploration of the simplex structure of the AHSS essay test in conjunction with the subscores of the AHSS objective test to determine whether the essay test demands more complex mental processes such as the ability to synthesize and organize arguments in a logical manner than are required by the subtests of the objective AHSS test.

The Sample

Of the 11,856 candidates who took the AHSS test in March 1971 only slightly over 2000 or about 1 in 6 also took the Verbal and Math aptitude as well as the English Composition Test. After identifying and matching the sample on all four tests, a final sample of 1977 candidates was used for the study. It was further observed that of all the 203,513 candidates in attendance at this administration only 13,716 or about 7% took both the aptitude tests and one or more achievement tests, with the remaining 93% taking either the aptitude or the achievement tests only. The sample for the study was, therefore, part of nearly 7% which had both achievement and aptitude test scores. To judge how typical of the total population the sample is, a comparison of the score statistics for each test is given in Table 1 for all candidates, and for the sample used for the study. Note that the total N of 11,587 for AHSS in Table 1 is the figure used for the final distribution of scores in the test-analysis report (Swineford, 1971) and differs from the total of 11,856 given above. This discrepancy may be accounted for by the answer sheets which were hand-scored and should scarcely have affected the total group score statistics in Table 1.

Insert Table 1 about here

As will be observed from Table 1 the matched sample of 1977 students used for the study obtained slightly lower AHSS scores than the total group. However, a comparison of this sample with the spaced sample of 915 selected for

test analysis purposes (with raw score mean of 31.68), which in turn is comparable to the total group mean, shows that both the study sample and the test analysis sample obtained an average weighted raw score of 15 on the AHSS essay test whereas the corresponding raw scores for the two samples on the AHSS objective test were 16 and 17 respectively. The conclusion from these results is that the slight difference between the study sample and the total group is in the objective test section.

The study sample also had a significantly higher mean score than the total group on SAT-verbal but a significantly lower score in the English Composition Test. There was no significant difference between the two groups in SAT-math.

The correlations of SAT-verbal and SAT-math with AHSS and the English Composition Test as well as the correlations between SAT-verbal and SAT-math are given in Table 2 below for the study sample of 1977 students and for the samples drawn from the entire group of 1968--69 and 1969-70 candidates for studying the scales for the achievement tests.

Insert Table 2 about here

In view of the closeness of the correlations for the three samples in Table 2, the moderately comparable score statistics in Table 1, and the large size of the matched sample selected for the study, the results of this investigation could be generalized to the entire American History and Social Studies population of the College Admissions Testing Program.

The Essay Test

When the College Board American History and Social Studies Committee met four years ago to discuss the possibility of including an essay as part of the achievement test in the subject it was decided that only one question should be set in order to minimize the scoring problems that were bound to arise if students were given a choice. It was further decided that the essay topic should not be just a "descriptive essay" but should be one that could also test skills such as synthesis, logical reasoning and the student's ability to engage in the process of historical analysis. The Committee also decided that the essay section should be allotted a testing period of 20 minutes out of the total testing time of 60 minutes fearing that a longer testing time might considerably lower the reliability of the objective section. With these as guidelines and with a further proviso that the topic should be one of sufficiently general interest as not to penalize some candidates unduly, four questions were assembled and pretested in January 1970. Routine ETS procedures were followed for the selection of schools for pretesting after insuring that SAT-verbal scores for potential schools were comparable to general candidate populations. The pretest population was found to have sufficient background on the American Revolution to answer the essay question on it; moreover, their essays were easily gradable and produced a good spread of scores.

Although the essay test was timed for 20 minutes (at the end of which candidates were instructed to proceed to the objective section), it was considered inappropriate to prevent any candidate who finished the objective

test in less than 40 minutes from returning to the essay since the total score on the test depended on his total performance during the 60-minute period. To monitor the candidates' reactions to the timing of the sections, a number of observers visited several test centers and reported that nearly all candidates used only the 20 minutes allotted to the essay, and started the objective section as soon as the supervisor instructed them accordingly. As it happened, most candidates could not finish the objective section within 40 minutes and therefore had no extra time to return to the essay. A few odd candidates who completed the objective test before the 60-minute period elapsed closed their test books and sat quietly for the remaining time. It is, therefore, safe to conclude from the observations that the timing of 20 minutes was strictly adhered to by the candidates.

Due to the fact that an issue of the College Board Review carried a report that the AHSS essay test would be given in March 1971, it was feared that the results of the study might be biassed by this announcement, especially if a large number of candidates who had prior knowledge of the essay made special preparation for it while an appreciable porportion had no such knowledge or preparation. To ascertain the extent to which the report had been widely circulated the AHSS answer sheet included a question requiring the candidates to indicate whether or not they knew of the essay section in advance. Out of a sample of 2099 answer sheets checked, only 26 candidates in 11 states indicated prior knowledge of the inclusion of an essay section. This number was considered too small to warrant a separate analysis.

Each essay script was scored by two different readers on an essentially four-point scale. The readers (who were assembled at one location for the scoring) were asked to assign grades on the basis of their overall impression rather than giving a fixed number of points to specific attributes. They were instructed to assign most papers to one of four categories, with scores of 2, 4, 6 or 8. Scores of 3, 5, and 7 were to be used for the relatively few papers that were difficult to classifv. Readers were also permitted to add or subtract a maximum of one point for unusually good or unusually poor expression. Thus, any one reader may assign scores ranging from the lowest score of 1 to the highest score of 9. An unscorable or blank paper received a total score of 1. Each candidate's total score is the sum of the scores assigned by the two readers, thus producing a range of raw scores of 1 to 18. Each of the 22 readers had special alphabetic codes for the grades he assigned to his papers. The codes for a particular reader were unknown to the others, so that the second reader could not decipher what grade had been assigned by his predecessor. Scoring of the papers was preceded by two days of standardization meetings between table or group leaders and the Chief Reader and by another day of coordination among the readers, their group leaders, and the Chief Reader.

The table or group leaders also checked random samples of each reader's papers at frequent intervals. Discrepancies of more than two points between the two readers of each essay were resolved by having the papers reread by the original readers. Where this process did not eliminate the discrepancy, the Chief Reader became the final arbiter.

The correlation between the scores assigned by the "first" readers and the "second" readers is .70 leading to an estimate of .82 for the reading reliability of scores obtained from the sum of two readings when the Spearman-Brown prophecy formula is applied (Swineford, 1971). The essay scores were later weighted before combining them with the scores on the objective section for a total score. The weighting for the total score was one part essay to two parts objective test thus weighting the two parts in proportion to the timing.

That the readers used the full range of the score scale can be seen by examining the frequency distribution of the sum of two readers' scores which is given below for the study sample of 1977 cases.

Score	f	Percent
17	12	0.6
16	71	3.6
15	4	0.2
14	152	7.7
13	14	0.7
12	451	22.8
11	15	0.8
10	415	21.0
9	16	0.8
8	481	24.3
7	8	0.4
6	181	9.2
5	7	0.4
4	121	6.1
1	29	1.5
Total	1977	
Mean		9.7
S.D.		3.1

The Variables Used for the Study

With the exception of the AHSS essay section, the four tests used for the study contained 5-choice objective test items. A more detailed description of the tests' contents is given on pages 16 to 18 below. The multiple-choice sections were formula-scored.

The 70-item objective AHSS test was quite difficult for the total group with more items answered wrong than right, and more than one-fifth not answered at all. This test appears to have been speeded for the group.

The score statistics for the matched sample of 1977 used for the study and the reliability estimates for the four tests are given in Table 3.

Insert Table 3 about here

For some of the analyses proposed in this study, the four tests were divided into subtests to yield total and subscores on 22 variables described below. The classification of the items was carried out by the Test Development Division at ETS in such a manner as to order the subtests of a particular subject in a hierarchical order of the mental processes involved in solving them (Bloom, 1956). With the exception of the scores for the 70 AHSS objective test items and the AHSS essay test which were standardized to a mean and standard deviation of 33.33 and 6.67, and 16.67 and 3.33 respectively, the raw scores for each of the other 20 variables were standardized to a mean of 50 and standard deviation of 10. A description of the 22 variables follows:

AdSS Variables:

1. AHE - The 20-minute essay test scored by two readers.

2. AHO - Comprising 70 objective items in American History and Social Studies.

3. AHT - The total AHSS test combined in the ratio of 1 part for the essay plus 2 parts for the objective test.

4. H01 - 29 AHSS objective items involving knowledge of specific terms, facts, conventions, concepts, principles, and theories.

5. H02 - 11 AHSS objective items involving comprehension of the meaning of a verbal or symbolic communication; ability to interpret or analyze graphic, pictorial or written material.

6. H03 - 8 AHSS objective items involving the extension of meaning beyond the content of a communication; ability to go beyond the given material to determine corollaries, implications, consequences; to make inferences, to make interpolations.

7. H04 - 22 AHSS objective items, 20 of these items involve the application of abstractions to particulars, the ability to select and apply hypotheses, concepts, theories or principles to given data; one involves the ability to judge the value of data for a given purpose; and the last involves the synthesis of given data into a new pattern.

SAT-verbal Variables:

8. VT1 - 18 Antonym items on the SAT-verbal test.

9. VT2 - 18 items of the sentence completion variety involving recognition of the logical and stylistic consistency among the elements of a sentence.

10. VT3 - 15 Verbal items on the comprehension of a prose passage.

11. VT4 - 34 Verbal items on inferring the specific relationships between words and ideas; drawing inference or making reasonable applications of the principles or opinions expressed in a passage.

12. VT5 - 5 Verbal items on evaluation of the author's logic, style and presentation.

13. VTT - Total of 90 SAT-V items (i.e., VT1 - VT5 above).

SAT-math Variables:

14. MT1 - 4 Math items on solving routine problems and performing mathematical manipulations.

15. MT2 - 25 Math items involving a demonstration of the comprehension of mathematical ideas and concepts.

16. MT3 - 11 items of nonroutine problems requiring insight or ingenuity.

17. MT4 - 20 Math items requiring the application of "higher" mental processes to mathematics. These include 18 items of the "Data Sufficiency" type.

18. MTT - Total of 60 SAT-M items (i.e., MT1 - MT4 above).

ECT Variables:

19. EC1 - 35 English Composition items on recognition of appropriate English usage.

20. EC2 - 35 items involving error recognition.

21. EC3 - 25 construction shift items.

22. ECT - Total of 95 items on the English Composition Test (i.e., EC1 - EC3 above).

Results

Correlations

The intercorrelations of the major variables used in the study are displayed in Table 4 for the sample of 1977 candidates.

Insert Table 4 about here

Comparing the three correlation coefficients between the essay, the AHSS objective and the combined essay and objective test, it is observed that the objective test contributes more to the total AHSS score than the essay as one would have expected. Furthermore, the moderate correlation of .47 between the essay and the objective test indicates that the proportion of common variance between the two sections is 22%. If both variables were measured with perfect reliability the correlation would be .77 thus yielding a common variance of 59% between them. The moderately low correlation between the two sections of the AHSS test suggests that they measure different psychological characteristics. Similarly, the low correlation of .401 between the obtained scores on the AHSS essay and the English Composition Test indicates that the former was measuring something other than writing skills since a correction for attenuation produced a correlation of .63 or a common variance of 39% between the two variables. As

had been observed in the preceding section, the correlations among SAT-verbal, SAT-math and English Composition are typical of similar results obtained in other administrations of the College Admissions Testing Program.

Further indication of the contribution of the AHSS essay test was provided by the correlation of the weighted composite score (C)--made up of AHSS objective test, SAT-verbal, SAT-math and English Composition--with the sum of C and the AHSS essay (E). That is, the correlation sought was: $r_{C(E+C)}$. Had the essay test ranked the candidates in the same order as the weighted composite score a perfect correlation would have resulted. Instead, a correlation coefficient of .997 which is hardly different from 1.00 was obtained, that is:

$$r_{(E+C)(C)} = .997$$

where

C = Sum of scores for AHSS objective (Variable 2), SAT-V (Variable 13), SAT-M (Variable 18) and ECT (Variable 22). The weights given to these variables in the composite were 2/3, 1, 1, and 1, respectively.

E = AHSS Essay (Variable 1), with a weight of 1/3 compared to 2/3 for AHSS objective.

The AHSS essay was, therefore, given a weight of 1/12 relative to the total weight for all four subjects.

However, suppose it were decided to vary the weight of the essay relative to that of the remaining tests used for the study in an attempt to maximize its contribution to the weighted composite score, then the effect[2] of the

various essay weights on the candidates' rank order (when the weighted essay scores are added to the weighted composite score made up of the AHSS objective, SAT-V, SAT-M and ECT) would be determined as follows.

Let $\underline{a}$ be the fractional weight to be given to the essay scores, E, relative to all four tests used for this study, and $\underline{b}$ the fractional weight assigned to the composite scores, C, in the remaining sections which make up the four tests such that

$$\underline{a} + \underline{b} = 1 .$$

Assume also that the scores, C and E, have been standardized so that $\sigma_C = \sigma_E = 1$.

$$r_{(aE+bC,bC)} = \frac{Cov(aE + bC,bC)}{\sigma_{(aE+bC)}\sigma_C}$$

$$= \frac{b + ar_{EC}}{\sqrt{a^2 + b^2 + 2abr_{EC}}} .$$

Substituting the correlation coefficient of .5057 for r_{EC} calculated the sample of 1977 candidates in the above formula and assigning various weights to $\underline{a}$ for the essay, the correlations obtained for $r_{(aE + bC,bC)}$ are given in Table 5.

Insert Table 5 about here

It is noticed from the results of Table 5 that only when the essay is assigned three-tenths of the overall weight of the combined total score in History,

SAT-V, SAT-M and ECT does its inclusion in the four-test battery change the original rank order by a noticeable amount.

If, therefore, the essay test is to make any difference in deciding who is to be selected for admission by altering the rank-order of the candidates even to the small extent indicated by a departure of the correlation coefficient from 1.00 to .96, it should be given a larger relative weight of probably no less than three-tenths of the total weight allotted to the four tests. A graph of the results tabulated in Table 5 is given in Figure 1.

Insert Figure 1 about here

Index of Uniqueness

An index of uniqueness, I_y^2, between a criterion variable y and n predictor variables $x_1, \ldots, x_n$ is a measure of the proportion of variance in y scores remaining, after taking into account the true-score correlation between y and the weighted composite of the predictors (Flanagan, 1959). This coefficient is determined from the following formula:

$$I_y^2 = r_{yy} - \frac{R^2_{y \cdot x_1 \ldots x_n}}{r_{cc}}$$

where

r_{yy} = reliability coefficient for the criterion variable y

R^2 = multiple correlation of variable y with variables $x_1, \ldots, x_n$

r_{cc} = reliability estimate of the weighted composite of the variables $x_1, \ldots, x_n$.

The uniqueness coefficient is, therefore, the proportion of the variance in y scores which is both reliable and completely free of overlap with any of the other tests.

The reliability estimate of the weighted composite of the independent variables (i.e., History Objective, SAT-V, SAT-M and ECT) was estimated as .9663, using the reliability estimates of each of these variables in Table 3. Applying this result and the multiple R of .526 to the above formula, the proportion of variance in the essay remaining after taking into account its true-score correlation with the weighted composite of the predictor variables is:

$$I_y^2 = r_{yy} - .286 = .45 - .286 = .\underline{164} .$$

The lower the reliability estimate of the essay, the smaller the uniqueness index. Thus, for a reliability coefficient of .40 which is not an unrealistic figure for a 20-minute essay scored by two readers (Coffman, 1966; Godshalk et al., 1966), the uniqueness index could be as low as .114. The relative magnitude of the obtained uniqueness coefficient for AHSS essay could be assessed by comparison with similar coefficients reported for the achievement tests given during the 1968-69 administration of the College Board admissions test (Coffman, 1971, p. 76). These coefficients (which represent the proportion of variance remaining after taking into account the true-score correlation of each achievement test with an optimally weighted composite of true scores on SAT-V and SAT-M) were as follows: .117 and .110 for English Composition and Literature respectively; .119 and .122 for Mathematics Level I and Level II; .318 and .323 for the one-hour objective test in American History and European

History; and .492 to .814 for Foreign Languages. Thus, the unique contribution of the AHSS essay in the present study is quite low, though comparable to the uniqueness coefficients obtained for English Composition and Literature tests, and Mathematics Level I and Level II.

Regression Analysis

In the absence of an external AHSS criterion by which the predictive validity of the AHSS essay, the AHSS objective test or a combination of both could be compared, a step-wise regression analysis was carried out to determine how well the essay scores could be predicted from a combination of the History Objective test, SAT-V, SAT-M and ECT. The variables were then entered into the regression equation, one at a time, according to their contribution to the multiple correlation. The results of the step-wise regression analysis showed that AHSS objective, SAT-M, SAT-V and ECT were entered in that order in the regression. The multiple correlation with the essay scores after the addition of each successive variable was .4725, .5149, .5256 and .5258.

Thus, the multiple correlation of the essay test with all four independent variables was .5258, with the percent of variance in the essay held in common with these four variables being 27.6. A large portion of the variance in the essay was, therefore, left unaccounted for; however, much of this could well have been the result of a prediction error due to the low reliability of the criterion. Using the estimate of .45 (Swineford, 1971, p. B) as the reliability coefficient of the criterion, the percent of reliable variance in the essay unaccounted for by the predictor variables is 17.4 (i.e., the difference between 45 and 27.6).

The prediction equation obtained from the multiple regression by using the standardized regression weights is as follows:

$$X_1 \text{ (Essay)} = .2720X_2 + .1687X_{18} + .1460X_{13} + .0263X_{22}$$

where X_2 = AHSS Objective test

X_{18} = SAT-math

X_{13} = SAT-verbal

X_{22} = English Composition .

Since all the predictor variables are correlated with the criterion their relative weights in the regression equation are fairly stable and not seriously affected by the low reliability of the criterion. As had been observed earlier, the exclusion of the English Composition Test from the regression model reduces the multiple R by .0002. The regression equation obtained after eliminating this test is:

$$X_1 = .271X_2 + .1733X_{18} + .1656X_{13} .$$

Discriminating Power of the Essay in Four Ability Groups

It should be recalled that one of the questions under investigation is whether the essay discriminates better among candidates of lower general ability than among those of higher ability.

The answer to this question is provided by the regression of essay on the Objective AHSS test in each of four groups, A, B, C and D, divided according to their scores on the combined Verbal and Mathematical aptitude tests. The

regression lines for predicting scores on the essay test from those of the objective test are drawn in Figure 2 for each of the groups. All candidates with scores above one standard deviation from the mean on SAT-V and SAT-M were assigned to Group A; those in the next lower category with scores above the mean were assigned to B; those in the third lower category with scores less than one standard deviation below the mean were in Group C; and the rest were assigned to D.

Insert Figure 2 about here

The slope of each regression line (given by the coefficient of the independent variable X) is a good indication of the extent to which the essay discriminates within each category--the steeper the slope, the greater the discrimination. It is obvious from the graph that a fixed difference between a pair of scores in each of the four groups on the objective test will result in disproportionate differences in corresponding essay scores, with the largest difference occurring in Group D. As may also be observed, Groups B and C have virtually parallel regression lines and discriminate equally whereas Group A with a slope of less than 1 in 10 exhibits the least score differences in the essay for corresponding changes in the objective section probably because of a ceiling effect in the regression.

Since the regression slope of the lowest ability group is the steepest, one may conclude that the essay test discriminates better among the candidates in that group than in any of the other three. However, the regression slopes are so closely parallel that the differences in the within-group discriminability are hardly of much real significance.

Factor Analysis

A factor-analytic study of the 17 independent subscores into which the four tests were divided was conducted by the Minimum Residuals method (Harman & Jones, 1966). This method was designed to minimize the residual correlations (i.e., the differences between the observed correlations and those reproduced from the factor analysis model). It yields a solution which best reproduces the original correlation matrix. The 17 variables used for the factor analysis have already been described in an earlier section and were included in the total of 22 variables detailed on pages 16 to 18. Those which are either total scores or a linear combination of any of the 17 variables were omitted. As already indicated, the subscores of each test were made up of items grouped together by ability dimensions according to an assumed hierarchical order of the complex mental processes involved in answering them (Bloom, 1956). Part of the rationale for this approach was to see if the abilities towards which the History essay was directed could be isolated as a unique factor apart from identifying any other variables with significant loadings on the same factor as the essay.

On the assumption that the cognitive ability dimensions tapped by the tests were no more than six, factor-analytic solutions for 2, 3, 4, 5 and 6 factors were computed.

The most satisfactory solution obtained was for four significant factors. These were later rotated to a varimax criterion. The rotated factor matrix of the four-factor solution is presented in Table 6.

Insert Table 6 about here

Since the communality of .999 shown in Table 6 for Variable No. 17 (VT4) is virtually a Heywood case and could indicate a linear dependency, the Minres solution which requires matrix inversion was compared with a steepest descent solution (Boldt, 1965) which does not. The methods produced nearly identical loadings. Moreover, the communality for variable MT4 was found to be 1.03 for the four-factor solution obtained by the second method. An interpretation of the factor analysis results, insofar as the essay is concerned, would hardly be changed even if the lower-bound reliability estimate (K-R#21) of .71 obtained for MT4 were substituted for its communality estimate.

The percent of variance in the factor space accounted for by the four factors was 63.3% leaving the difference of 36.7% as unique and/or error variance. The highest loading of .4151 for the essay test appeared in the same factor on which the four subscores of the AHSS Objective test had the highest loadings. The proportion of the essay test's variance accounted for by this factor was 17% out of the total of approximately 30% of its total variance involved in the factor space. The lowest communality of .297 obtained for the essay suggests that it measures one or more psychological factors different from any inherent in the other 16 variables. Thus, the proportion of reliable variance in the essay not held in common with the other variables was found to be .15. This proportion represents the difference between the reliability and the communality estimates--a result consistent with those of the uniqueness and the regression analyses. The fact that the essay test had no

significant loadings (arbitrarily set as .30 or more) on any of the other factors particularly those on which the Verbal and English Composition subtests had the highest loadings again confirm that the essay was essentially not a test of verbal ability nor could it be described as a reasoning test since it also failed to cluster with the factor which was highly loaded with the Mathematics subscores.

Discriminant Function Analysis

The discriminant function analysis--a statistical method of combining test scores so as to maximize the differences between groups of persons and minimize the differences within each group--was used to determine the relative weights of a number of independent variables in characterizing multivariate differences among five ability levels or groups.

A computer program for multivariate analysis of variance (Finn, 1968) which also incorporates a routine for discriminant function analysis was used for the study. The analysis was designed to compare the performance of five groups divided according to their quintile ranks on SAT-verbal and math combined. As expected, highly significant differences between the group mean vectors on the independent variables were obtained, but these results are of little special interest to the study and will, therefore, not be displayed.

After dividing the sample into the five ability categories specified above, three independent variables (i.e., AHSS essay, AHSS objective test and ECT) hypothesized to be related to membership in the five ability groups in view of their use for supplementing the aptitude scores in the college admissions process, were entered into the discriminant function analysis in order to determine the contribution of each variable in separating the five groups.

The standardized coefficients of the "best" linear function of the three variables which distinguishes (or "discriminates") among the quintile groups are as follows:

Variable	Standardized Discriminant Function Coefficient
1. AHSS Essay	0.2001
2. AHSS Objective	0.4430
22. English Composition	0.7840

The percentage of between-group variation accounted for by the new variate:

$$Y = .2001\ X_{\uparrow 1} + .4430\ X_{\uparrow 2} + .7840\ X_{\uparrow 22}$$

(Essay) (Objective AHSS) (ECT)

was found to be 98.97. This implies that a combination of these three variables in the given proportions will virtually reproduce the original grouping of the sample into the five ability categories. The chi square value for testing the significance of this variate was 2236 with 12 degrees of freedom which is highly significant at p less than .0001.

An important aspect of this result is that the essay section contributed slightly less than one-half the weight of the objective section to the group separation. In interpreting this result it should be recalled that the standardized discriminant function coefficients indicate the relative contribution of each variable after entering them into the analysis with equalized variances.

Simplex Structure

A test of hypothesis that an essay test demands more complex behaviors (such as organization and synthesis of ideas in an original production) than are required for a multiple-choice test was investigated by analyzing the correlation matrix between ordered subtests of the AHSS test for simplex structure. The typical property of a simplex correlation structure is that correlations for an ordered series of subtests decrease as one moves away from the main diagonal. This pattern should, therefore, be observed in the data, with the essay subtest (presumed to be the most complex) having a higher correlation with HO4 (Variable 7) than with HO3 (Variable 6), and decreasingly with HO2 (Variable 5) and HO1 (Variable 4) in that order.

The correlation matrix for the five variables shown in Table 7 was analyzed for simplex structure by a generalized computer program (Jöreskog, 1970; Jöreskog, Gruvaeus, & van Thillo, 1970) for the estimation and testing of simplex models.

Insert Table 7 about here

The result of the analysis was inconclusive and showed that either the variables were not in the correct order or the model for testing the simplex structure did not fit the data. The inconclusive nature of this result is probably due to two reasons: (1) either some of the objective test items may have been misclassified or (2) the scaling of the subtests does not hold for the subject since the underlying assumption in fitting a set of data to a simplex model is that a person who answers a more difficult or complex question

satisfactorily should also answer less complex ones correctly. A third possible reason could probably hinge on the speededness of the objective section which obviously precluded many candidates from attempting a good number of items towards the end of the test. Furthermore, the implicit assumption that all the candidates had covered the content specifications of the test well enough for the scores to reflect their ability levels could be false. That the essay correlates more with HO1 than any of the other subtests suggests that the essay scores may have reflected the factual content of the answers more than the other skills.

The KR#21 reliability estimates of the four subtests: HO1, HO2, HO3 and HO4 were found to be .658, .448, .166 and .557 respectively. If the correlations of the essay with the four subtests, HO1 to HO4 were corrected for attenuation, the results in the last row of Table 7 would become .537, .562, .697 and .517 which, again, do not form an ordered series as had been expected. An attempt to correct all the correlations in Table 7 for attenuation, using the KR#21 reliability estimates for HO1 to HO4, produced some coefficients greater than unity which suggests that the reliabilities may have been underestimated. Any interpretation of the corrected correlation coefficients should, therefore, be viewed with caution.

Conclusions

The results of the several independent statistical analyses conducted for the study support the conclusion that insofar as the pool of information available from the scores in the objective section of the American History test, the SAT-verbal, SAT-math and English Composition tests is concerned, the additional information gained by the introduction of an essay section is minimal.

This conclusion is supported by the following results:

(1) The correlation of the composite score (C)--made up of AHSS objective test, SAT-verbal, SAT-math and English Composition weighted in the proportion 2, 3, 3, 3 respectively--with the sum of C and the AHSS essay (E) was found to be .997. Had the essay section ranked the candidates in the same order as C, a perfect correlation of 1.000 would have resulted. Instead, a correlation coefficient for $r_{C(E+C)}$ of .997 which is hardly different from 1.00 revealed that the inclusion of the essay scores scarcely made any difference to the original ranking of the candidates based on the composite score, C. Even if the weight of the essay were increased to three-tenths of the entire four-test battery, the correlation of .997 would drop further to only .9568, so that, for all practical purposes, virtually the same students are likely to be accepted on the basis of their scores in the battery, either with the three-tenths weighting for the essay or without it altogether. It becomes ever less important when other factors such as high school grades enter into the college admissions decision process. Considering the total AHSS test by itself, it should be noted that the moderate correlation of .47 between the obtained essay and the objective test scores suggests that they measure different psychological characteristics. The upper limit of the correlation between the scores on the two sections, after correcting for attenuation, is .77 which again confirms the conclusion that they measure different psychological attributes.

(2) An attempt to predict the essay score from the scores on AHSS objective section, SAT-verbal, SAT-mathematical and English Composition showed that the proportion of variance in the essay scores accounted for by these

predictor variables was .28 out of a maximum possible common variance of .45 as indicated by the estimated reliability of the essay. Thus, only .17 of the remaining variance in the essay was left unexplained by the four predictor variables.

(3) The result of the regression analysis was also confirmed by the uniqueness analysis from which the proportion of variance in the essay that is both reliable and completely free of overlap with a weighted composite of the four-test battery (excluding the essay) was found to be .16. However, it should also be noted that this analysis resulted in about as much unique variance as was produced for the English Composition Test after its correlation with an optimally weighted composite of SAT-verbal and SAT-mathematical (Coffman, 1971). Again, considering the American History and Social Studies test by itself, it was found that the proportion of variance remaining in the essay scores--after taking into account their true-score correlation with the scores on the objective section--was found to be approximately 22%.

(4) The factor analysis of the 17 subtests into which the four tests (American History and Social Studies, SAT-verbal, SAT-mathematical and English Composition) used for the study were classified by ability dimensions identified four significant factors which accounted for 63.3% of the total variance in the common factor space. Of this total variance, the essay accounted for only 2.7%. It also had the least communality which suggests that it measures one or more psychological factors different from those inherent in the other 16 variables. However, much of this specific factor could be attributed tc error variance, leaving only a reliable variance unique to the essay of .15. The essay subtest also had a significant loading (arbitrarily set at .30 or higher) of .4151 on

only one factor on which all the American History objective subtests had the highest loadings. Thus, the essay was found to be essentially a test of attainment in American History and Social Studies contrary to the fears expressed by some of its opponents that it would measure primarily verbal skills and literary quality.

(5) The intercorrelation of the essay with the four subtests into which the objective AHSS test was subdivided (according to a hypothesized order of complexity of the cognitive ability they are presumed to assess) indicated that the essay was more analogous to the subtest that demanded mere factual recall than to that which measured the ability to select and apply hypotheses, to synthesize and to judge the value of data for a given purpose, etc. Thus, the results for this one essay failed to support the notion that essays demand more complex mental processes than objective tests.

(6) The discriminant function analysis--a statistical method of determining the combining weights for a number of variables in order to "discriminate" or "distinguish" maximally between groups or classes of persons--was used to determine the relative weights for a linear combination of AHSS essay, AHSS objective and English Composition which could best distinguish between groups of candidates divided into five ability categories according to their composite scores on SAT-verbal and SAT-math. The combining weights which also indicate the relative importance of the contribution of each variable in distinguishing between the five groups were: .2001 for AHSS essay, .4430 for AHSS objective and .7840 for English Composition. These weights again confirm the conclusion that the effectiveness of the essay, even as part of a pool of three variables with equal variance, is minimal in discriminating between the five ability groups.

References

Bloom, B. S. (ed.) Taxonomy of Educational Objectives: Handbook I, Cognitive Domain. New York: David McKay Co. Inc., 1956.

Boldt, R. F. Factoring to fit off-diagonals. Technical Research Note 161. Washington, D. C.: U. S. Army Personnel Research Office, 1965.

Coffman, W. E. On the validity of essay tests of achievement. Research Memorandum 66-2. Princeton, N. J.: Educational Testing Service, 1966.

Coffman, W. E. The achievement tests. In W. H. Angoff (ed.), The College Board Admissions Testing Program. New York: College Entrance Examination Board, 1971.

Finn, J. D. Multivariance--Univariate and Multivariate Analysis of Variance and Covariance: A FORTRAN IV Program. (Version 4) Buffalo, N. Y.: State University of New York at Buffalo, 1968.

Flanagan, J. C. Flanagan Aptitude Classification Tests: Technical Report. Chicago, Ill.: Science Research Associates, 1959.

Godshalk, F. I., Swineford, F., & Coffman, W. E. The Measurement of Writing Ability. New York: College Entrance Examination Board, 1966.

Harman, H. H., & Jones, W. H. Factor analysis by minimizing residuals (MINRES). Psychometrika, 1966, 31, 351-368.

Huddleston, E. M. Measurement of writing ability at the college level: Objective vs. subjective testing techniques. Journal of Experimental Education, 1954, XXII, 165-213.

Jöreskog, K. G. Estimation and testing of simplex models. Research Bulletin 70-42. Princeton, N. J.: Educational Testing Service, 1970.

Jöreskog, K. G., Gruvaeus, G. T., & van Thillo, M. A general computer program for analysis of covariance structures. Research Bulletin 70-15. Princeton, N. J.: Educational Testing Service, 1970.

Swineford, F. College Entrance Examination Board Achievement Examinations, March 1971. Statistical Report 71-73. Princeton, N. J.: Educational Testing Service, 1971.

Acknowledgements

[1]The author would like to express his appreciation to Robert F. Boldt and John Fremer of Educational Testing Service for their comments as reviewers.

[2]Grateful acknowledgement is also due to Edward O'Connor of Educational Testing Service who suggested this approach to the study.

Table 1

Means and S.D.'s for the Total Group for Each Test Compared to Those for the Study Sample

		AHSS Total		SAT-V		SAT-M		ECT	
		Raw	Scaled*	Raw	Scaled	Raw	Scaled	Raw	Scaled
Mean	Total Group	31.91	463	34.10	446	23.40	474	37.70	497
	Study Sample	30.49	452	37.34	466	23.43	474	32.68	466
S.D.	Total Group	13.10	101	17.11	109	12.72	113	17.08	106
	Study Sample	12.01	93	17.30	110	12.16	108**	17.41	108
No. of Cases	Total Group	11587		170533		170521		35268	
	Study Sample	**1977**		**1977**		1977		1977	

*College Board Scale of 200-800.

**Estimated.

Table 2

Correlations among SAT and Achievement Tests

	Am. History & Soc. Studies			English Composition		
	1971 Study Sample	1968-69 Sample	1969-70 Sample	1971 Study Sample	1968-69 Sample	1969-70 Sample
No. of Candidates	1977	4634	2901	1977	4701	3552
Correlation						
SAT-V vs. Ach. Test	.71	.72	.71	.82	.81	.78
SAT-M vs. Ach. Test	.54	.57	.56	.62	.60	.61
SAT-V vs. SAT-M	.63	.62	.63	.63	.60	.61

Table 3

Score Statistics, Range of Scores and Reliability Estimates

Test	Raw Score		Raw Score Range		K-R#20
	M	SD	Possible	Observed	Reliability Estimates
Am. History & S.S. (Total)	30.5	13.1	-15 to 97	- 6 to 82	.830
Essay (Unweighted)	9.7**	3.1	1 to 18	1 to 17	(. 45)*
Objective	15.8	10.3	-17 to 70	- 8 to 54	.832
English Composition	32.7	17.4	-24 to 95	- 6 to 87	.913
SAT-math	23.4	12.2	-15 to 60	- 5 to 56	.911
SAT-verbal	37.3	17.3	-22 to 90	- 9 to 86	.928

*Estimated Test Reliability which reflects the error associated with the sampling of questions and the error associated with the unreliability of reading.
This coefficient was estimated from the results of previous studies.

**This was converted to a weighted raw score mean of 14.7 before combining it with the raw score on the objective section to ensure that the weighting of the essay to the objective section is in the proportion of 1:2.

Table 4

Table of Intercorrelations

		1	2	3	13	18	22
1.	AHSS (Essay)						
2.	AHSS (Objective)	.47	---				
3.	AHSS (Essay & Objective)	.74	.94	---			
13.	SAT-verbal	.46	.70	.71	---		
18.	SAT-math	.41	.50	.54	.63	---	
22.	English Composition	.40	.55	.58	.82	.62	---

Table 5

Effect of Changes in Essay Weights on Correlations

Essay Weight (a)	Composite Score Weight (b)	$r_{(aE+bC,C)}$
0	1	1.000
1/12	11/12	.9972
1/10	9/10	.9959
2/10	8/10	.994
3/10	7/10	.9568
4/10	6/10	.9186
5/10	5/10	.8676
9/10	1/10	.5816
11/12	1/12	.5688

Table 6

Factor Analysis Results

	Variable		Rotated Factors** 1	2	3	4	Communality h^2
AHSS Obj. Subtests	4.	HO1	.2253	.1932	.7513*	.1349	.671
	5.	HO2	.2176	.1519	.6858*	.1137	.554
	6.	HO3	.1365	.1112	.5491*	.1136	.345
	7.	HO4	.2632	.2118	.6466*	.1576	.557
ESSAY	1.	AHE	.2444	.2477	.4151*	.0636	.297
VERBAL Subtests	8.	VT1	.5567*	.2034	.3950*	.4117*	.677
	9.	VT2	.5820*	.2410	.3886*	.4762*	.775
	10.	VT3	.5557*	.2679	.4256*	.2531	.626
	11.	VT4	.5905*	.3413*	.4603*	.3039*	.769
	12.	VT5	.3488*	.1098	.2489	.6330*	.596
MATH Subtests	14.	MT1	.1567	.6046*	.1530	.0845	.421
	15.	MT2	.2830	.7932*	.2514	.0765	.778
	16.	MT3	.2673	.6770*	.1898	.1183	.580
	17.	MT4	.2177	.9534*	.1922	.0767	.999
ECT Subtests	19.	EC1	.7432*	.2952	.2251	.1581	.715
	20.	EC2	.7431*	.2623	.2813	.1582	.725
	21.	EC3	.7091*	.3097*	.2473	.1338	.678
							10.763

*Factor loadings of .30 or more.

**These factors may be labelled as (1) English Composition/Verbal. (2) Mathematical Reasoning, (3) Historical Content/Verbal Reasoning and (4) Verbal Fluency.

Table 7

Correlation Matrix for Simplex Structure

	4	5	6	7	1
4. HO1	1.00				
5. HO2	.60	1.00			
6. HO3	.48	.46	1.00		
7. HO4	.62	.55	.42	1.00	
1. AH Essay	.44	.38	.28	.39	1.00

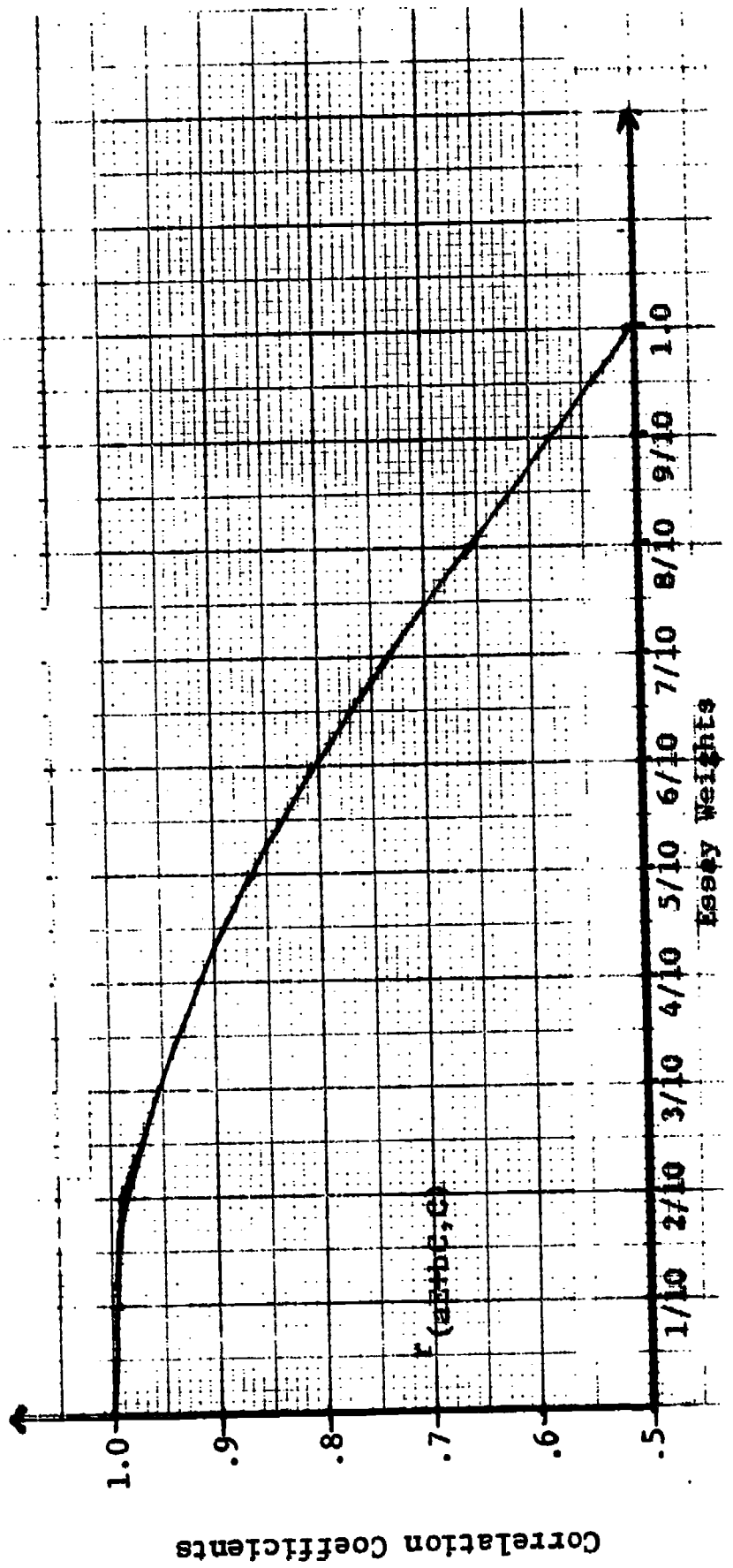

Figure 1: Effect of Changes in Essay Weights on Correlations

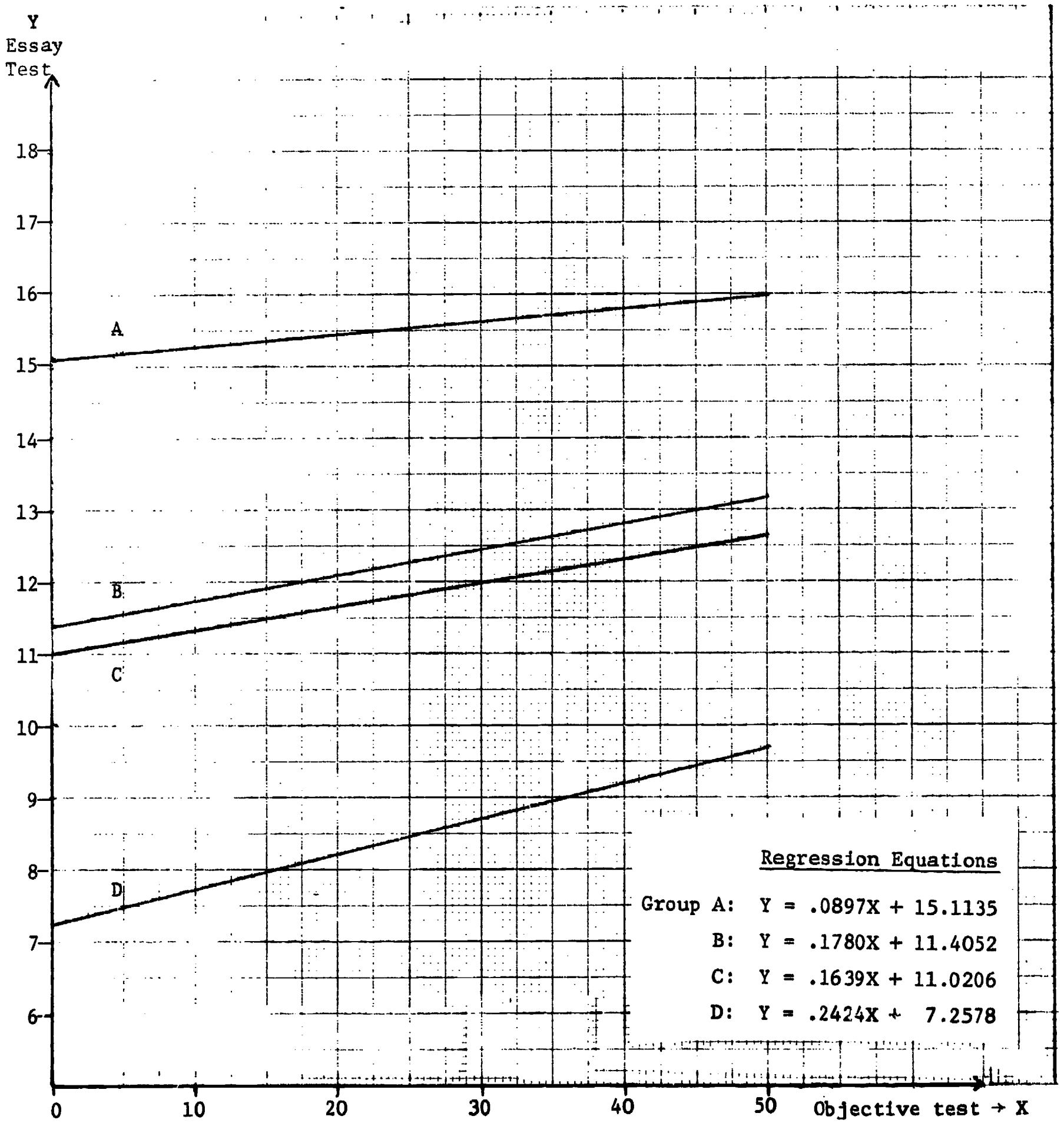

Figure 2: Regression of AHSS Essay on Objective Test in 4 Groups